2011 IEEE International Conference on Robotics and Automation
Shanghai International Conference Center
May 9-13, 2011, Shanghai, China

Learning Speaker Recognition Models through Human-Robot Interaction

E. Martinson, W. Lawson

Abstract—**Person identification is the problem of identifying an individual that a computer system is seeing, hearing, etc. Typically this is accomplished using models of the individual. Over time, however, people change. Unless the models stored by the robot change with them, those models will became less and less reliable over time. This work explores automatic updating of person identification models in the domain of speaker recognition. By fusing together tracking and recognition systems from both visual and auditory perceptual modalities, the robot can robustly identify people during continuous interactions and update its models in real-time, improving rates of speaker classification.**

I. INTRODUCTION

Person identification is the problem of identifying an individual uniquely by their observable characteristics. A few of the characteristics that successfully been deployed include face recognition [1], passphrases [2], speech characteristics [3], gait [4], and clothing [5]. Each works, to a varying degree of accuracy, by matching what is sensed by the computer system with a priori knowledge of the individual. But from where does the computer system acquire such knowledge? Even if the system has such a priori knowledge, models can change over time. Once the system has been deployed, models may need to be updated periodically to match changes in physical appearance and behavior. For use with a robot, however, this is a serious limitation. The robot cannot every day make active requests for the same information. This would lead to disuse in the long run as the robot's novelty wears off and updating becomes irritating. Instead, robots need to do what people do, and update models in real time to reflect the present.

The challenges with real-time updating are twofold. First, the robot needs to recognize the individual correctly with the current model before it can make necessary updates. But if the information is outdated, how can it guarantee accurate recognition? Secondly, the robot needs to continuously recognize the individual throughout the interaction. With the potential for multiple individuals during an interaction, the robot cannot rely on a single positive identity result. What if the first score was wrong or the people move about the room? People solve these problems by integrating a variety of perceptual modalities and tracking multiple characteristics simultaneously. A robot can do the same. In this work, our humanoid robot combines visual and auditory features to track and identify one or more individuals continuously during an interaction. Positive identifications, in turn, allow the robot to update its own speaker recognition models, improving classification rates as part of a virtuous cycle.

The real robot implementation of this work integrates 3 different perceptual systems: 1) RGB camera, 2) SR3000 depth camera, and 3) 4-element microphone array. The cameras enable visual face detection, recognition and tracking. The microphones then add speaker localization and recognition, letting the robot to rotate to face new speakers and better integrating incoming speech with the correct visual information. The combined system allows the robot to track a person during an interaction, fusing evidence over time to recognize the individual and begin updating models.

The remainder of this work is organized as follows: related work in fusing perceptual information is presented in the next section; section III describes individual recognition and tracking systems; section IV covers the combined implementation for real-time updating; and, finally, section V evaluates performance.

II. RELATED WORK

Many recognizable characteristics exist for use in person identification. Each characteristic excels in some domains and suffers in others. Visual characteristics, for instance, are affected by lighting conditions and viewing angle. Similarly, audio suffers in loud environments. For that reason, much recent work in person identification has focused on fusing different perceptual features and/or modalities to overcome environmental weaknesses. The specific features to fuse, however, vary widely between applications.

Face and speaker recognition are probably the most common fusion options. Palanivel and Yegnanarayana [6] combined these two modalities to recognize news anchors from television broadcasts. They combined information from these two recognition systems plus a tracker using a simple

Manuscript received Feb 8, 2010. This work was supported by the Office of Naval Research under job order numbers N0001408WX30007 and N0001410WX20773.

E. Martinson is a post-doctoral fellow with the U.S. Naval Research Laboratory, Washington, DC 20375. (phone: 202-404-4948; e-mail: eric.martinson.ctr@nrl.navy.mil).

W. Lawson is also with the U.S. Naval Research Laboratory, Washington DC (e-mail: wallace.lawson@nrl.navy.mil).

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Learning Speaker Recognition Models through Human-Robot Interaction	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Naval Research Laboratory,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Person identification is the problem of identifying an individual that a computer system is seeing, hearing, etc. Typically this is accomplished using models of the individual. Over time, however, people change. Unless the models stored by the robot change with them, those models will became less and less reliable over time. This work explores automatic updating of person identification models in the domain of speaker recognition. By fusing together tracking and recognition systems from both visual and auditory perceptual modalities the robot can robustly identify people during continuous interactions and update its models in real-time, improving rates of speaker classification.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**6**	

weighting rule for the actual fusion. The combined system showed a natural improvement over any individual modality. Because their data set is comprised of news broadcasts, however, the system does not deal with common robotic problems such as not facing the camera, or poor SNR. Martinson et al [7] demonstrated a similar improvement on a real robot, fusing face and speaker recognition together to overcome gaps in either face or speech detection to improve the regularity of recognition.

Salah et al [8] also integrated tracking, face recognition and speaker recognition together, this time for simultaneous identification of multiple people in a "smart" environment. They overcame poor visual angles and SNR problems by continuously tracking users from the moment they entered the room. Unlike robots, which often have a limited field of view, microphones and cameras covering all regions meant that users were rarely lost, and recognition scores could be combined over lengthy periods of time using a particle filter.

In each of the previously mentioned works [6,7,8] people are enrolled into the system prior to recognition. But others have begun to address the problem of missing training information. Belotto and Hu [5] use online training with a service-type robot to enroll users. Their robot uses a combination of clothes, height, and face recognition to identify enrolled individuals and follow them through an environment filled with unknown people. Other, more preliminary work, by Roy and Marcel [9], explores the reconstruction of missing audio/video recognition models from different perceptual modalities. For example, if a speaker is known only by voice, they could be recognized from lip movements in a video. Recognition precision suffers significantly, but even weak models could aid recognition.

III. Recognition and Tracking

This current work focuses on the problem of improving speaker recognition models through interaction with a humanoid robot. Because this is a real system, individuals do not always talk loudly or cleanly, and/or maintain eye contact with the camera. The result is a drop in recognition recall as likelihood scores remain low, meaning that a lot of possible training data is inappropriately discarded. The solution to this problem, as evidenced by [5,8], is to combine not just multiple recognition modalities, but also tracking in each modality. Our humanoid robot platform, the Mobile-Dexterous-Social (MDS) robot[1] called Octavia, enables face and speaker recognition, as well as person tracking and sound localization. The following section describes each of these capabilities as implemented on Octavia (Figure 1).

A. Face Recognition

Approaches to face recognition can be classified as either *closed-world* or *open-world*. In closed world approaches, the robot assumes that everyone is known and has been previously enrolled into a biometrics database. When encountering a new person, the acquired (probe) face will be matched up against the known (gallery) faces using a similarity measure. Gallery faces are ranked in terms of how well they match the person in question. The person with the best matching score is recognized. The difficulty with closed-world approaches is that we cannot assume a mobile-robot will only interact with a small set of known people. Closed world approaches can be modified by using a threshold to determine if the individual match is sufficiently close, but choosing a good threshold is challenging.

Fig 1. Octavia is an MDS robot with a 4-element microphone array mounted on the body, cameras in the eyes, and an SR3000 in the forehead.

Open-world approaches to face recognition are designed to recognize a small number of known individuals, while marking everybody else as unrecognized. We use the face recognizer developed by Kamgar-Parsi et al. [10]. It is based on identifying and enclosing the region R_T in the human face space that belongs to the target person T. If a face is projected inside R_T, it is identified as the target person. This approach is computationally efficient, appropriate for real-time recognition on a mobile robot, and has been extensively tested on rejection of unknown individuals while still showing a good ability to recognize a small set of known people. For more details, we refer any interested readers to the referenced publication.

In this work, subjects participated in two sessions each. The first session was used to train the network; the second session was used to test the network. Figure 2 shows the ROC curves evaluating our approach. The human study in which data was collected is described in more detail in [7].

B. Person Tracking

The purpose of our person tracker [11] is to detect new individuals in the scene as well as to track their face over time. When a face is detected, a new track is initialized with a face location and a skin color histogram. Depth information from the SR3100 time of flight camera is

[1] http://robotic.media.mit.edu/projects/robots/mds/overview/overview.html

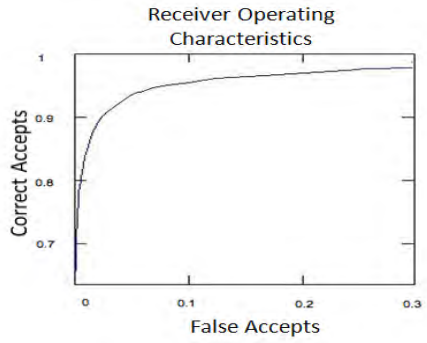

Fig 2. ROC curve showing the performance of our approach to face recognition. The y-axis shows the hit rate, while the x-axis shows the false accept rate

utilized in conjunction with the color histogram to determine whether or not the individual is still within view of the camera. Multiple faces can be tracked simultaneously by the same pair of sensors using this approach.

C. Speaker Recognition

Speaker recognition on Octavia is based on Gaussian mixture models (GMM), as described by Quatiri [3] for use in speaker verification. Models are created from non-overlapping audio samples collected by the fused-recognition system (Section IV) which have been labeled as belonging to a specific speaker. To build a speaker model, the first 10 mel-frequency cepstral coefficients (MFCC's) are extracted from each 10-msec frame in the recorded audio. The first MFCC is correlated to the energy of the audio segment, so it is used to separate speech from ambient noise. This speech detection method assumes that all loud, wide spectrum sounds heard by the robot are actually speech, but is otherwise effective for removing quiet frames. A GMM with 50 components is then created from all remaining MFCC speech vectors using k-means. The resulting model consists of a centroid μ_i, a covariance matrix σ_i, and a prior probability p_i for each component i. Except to identify the presence of speech, the first two MFCC's are not used in model creation, leaving only 8 dimensions ($R = 8$).

Models are created both for the speaker and an "imposter" created from all speech frames for other speakers in the data set. Given a speaker/imposter pair $\{S_k, I_k\}$, which will henceforth be called a speaker verifier, the odds of an audio segment belonging to speaker k are calculated as follows. First, MFCC vectors x_m are extracted for the entire segment and quiet frames are removed. Next, for each vector, the probability of belonging to either model M is determined:

$$p(\vec{x} \mid M) = \sum_{i=1}^{50} \frac{p_i}{\sqrt{(2\pi)^R |\sigma_i|}} e^{(\vec{x}-\mu)^T \sigma_i^{-1} (\vec{x}-\mu)} \qquad (1)$$

To avoid zeros in subsequent calculations, it is assumed that each vector had to belong to either the speaker or the imposter, and that each were equally likely. So, given an audio stream segment α_t, which contains a set of speech frames m, the odds of a speaker being present at time t are represented by Eq 2:

$$O_k(t) = O_k(t-1) + \sum_m \log\left(\frac{p(\vec{x} \mid S_k)}{p(\vec{x}_m \mid I_k)} \right), \forall m \in \alpha_t \qquad (2)$$

For use with continuous streams, O_k is restricted to the range [-40,40] preventing extremely large values. A decay function is also added to bring all speaker likelihoods back to neutral (i.e. 50% or O_k=0) when no one is talking. Whenever an audio segment contains no speech, O_k is reduced by -0.2*O_k across all speakers k.

When fully trained, speaker recognition performance varies with the number of active verifiers in the set. Figure 3 plots recognition of 2-sec speech fragments. With only 2 speakers active, precision is >90% and recall (percentage of correctly identified samples), is near 80%. With larger numbers of active verifiers, however, both drop significantly. How, therefore, can the robot recognize more people using speech? As described in [7], a solution to this problem is to use the results of face recognition to load the set of active speaker verifiers dynamically. When a speaker's face recognition likelihood surpasses a minimum score (usually much less than needed for an accept), their speaker model is loaded into the set of active speakers, and imposter models for all speakers are recalculated. Then, if there is no activity recorded from that speaker for a while, they are removed again from the active set. By keeping the set of active verifiers small, speaker recognition performance is elevated.

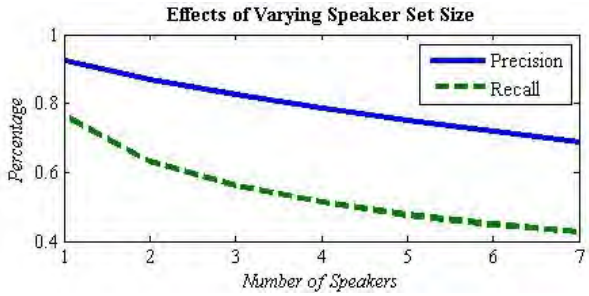

Fig 3. Precision and classification rate decrease with the number of speakers in the set.

D. Speech Localization

Speech localization on Octavia is determined by auditory evidence grids [12]. Every quarter-second sample of audio data from a four microphone array is evaluated for the presence of speech using the MFCC-based method described previously. Each sample is analyzed using generalized cross correlation to estimate energy from a hypothetical set of sound sources located 1-m from the robot, creating a spatial likelihood. The last 2-sec of recorded spatial likelihoods are then combined using log-odds notation to filter abnormal individual readings. When one strong, or multiple weaker measurements from a speaker create enough evidence (e.g. a threshold), a direction is assigned to the most recent sample. As with speaker recognition, samples not containing speech should lessen the likelihood of a person being present. This works by introducing neutral likelihoods, where all locations are equally likely for all non-speech samples.

IV. Learning Speaker Models

In the current implementation, person tracking, face recognition, and sound localization/speaker recognition are run asynchronously as separate services. Every 0.25s, the systems are polled to create an identification packet, *P*:

$$Track \in \left[T_{exists}, T_{direction}, T_{depth}\right]$$

$$Face \in \left[F_{location,1}, F_{prob,1}, \ldots, F_{location,n}, F_{prob,n}\right]$$

$$Audio \in \left[A_{exists}, A_{prob}, A_{direction}\right]$$

$$Packet \in \left[Track, Face, Audio\right]$$

Directions in this packet are in radians, and depths are in meters. "Prob" variables are arrays of likelihoods, one value per known individual, calculated by either face recognition or speaker recognition. Note that face recognition may detect multiple faces in the picture simultaneously, so a packet may contain multiple face recognition results.

These synchronized packets then form the basis for fusing the data and continuously recognizing the individual in order to update auditory models.

A. Perceptual Fusion

To combine the data, separate identity tracks are currently maintained on the continuous recognition system take 5 sec of inactivity to expire. These separate tracks are necessary in the current implementation because the person tracker depends on skin tone. Person tracks disappear when skin colored objects disappear from sight, only to be created anew with the next visible face. An identity track, therefore, is necessary to preserve history as people shift around:

$$People_i \in \begin{bmatrix} P_{direction,i}, P_{depth,i}, P_{prob,i} \\ P_{faces,i}, P_{bestID,i}, P_{best\_prob,i} \end{bmatrix}$$

Identity tracks are updated based on the perceived location of the face and/or audio results. When face recognition is analyzes a face within 0.3-rad of a known identity track, that track is updated. Similarly, when sound localization detects speech activity within 0.3-rad of a known identity track, that track is updated. The position of that track relative to the robot, however, is only updated by the person tracking system, which is the most accurate at localization. Identity tracks are only created when a new individual is tracked, and that object track's location and distance are within 0.3-rad, 0.5-m of an existing, unexpired track. The addition of any new data to an identity track keeps it from expiring.

The fusion rule for combining likelihoods is a weighted summation of the latest data assigned to a given track (Eq 3):

$$P_{prob,i}(t) = \frac{\left(W_h P_{prob,i}(t-1) + W_f F_{combined,i} + W_a A_{prob,i}\right)}{\left(W_h + W_f + W_a\right)} \quad (3)$$

The weights (W_h,W_f,W_a) were assigned to be (1.5,1.5,1.0). When data are missing, however, the weights are changed to zero for that data. Therefore, if no face is detected in this sample, or there are no new data from face recognition, then that W_f=0. Similarly, speaker recognition results are not included when no speech is detected. $F_{combined,i}$ is a variable created from the last three faces assigned to identity track, *i*. It is normalized only when its summation is >1 so that small face recognition scores do not bias the result (Eq 4). The same procedure is not repeated for speaker recognition because speaker scores are already updated on a continuous basis, preserving history between samples.

$$K_i = P_{faces,i}(t-1) + P_{faces,i}(t-2) + F_{prob,i}$$

$$F_{combined,i} = \begin{cases} K_i, & if \sum_j K_j \le 1.0 \\ K_i \Big/ \sum_j K_j & else \end{cases} \quad (4)$$

B. Learning through Continuous Recognition

Tracks provide the basis for continuous recognition by preserving the history of sensor-based recognition. Because the sensor readings are assigned to an updated location, as reported by the person tracker, speakers do not need to remain still or always face the robot either, actions which are unnatural for real interaction. But the fused recognition history alone does not confirm the identity of the speaker. Recognition results are still biased by the latest sensor readings, which may fluctuate substantially (see Figure 4). To collect enough data for speaker recognition, however, continuous recognition is required. Each track must preserve its identity between samples to smooth out variations.

A straightforward approach to assigning a track identity is to select the speaker with the highest recorded likelihood. The problem with this approach is that it assumes any errors made in recognition were of low magnitude relative to recognition scores for the correct individual. Unfortunately, neither face nor speaker recognition fail that way. Instead, both are right most of the time, but may fail spectacularly for one or two samples. This is particularly bad at the beginning of an interaction when people may be looking in a different direction, and there is not enough data to suppress bad face recognition results through filtering and perceptual fusion. A

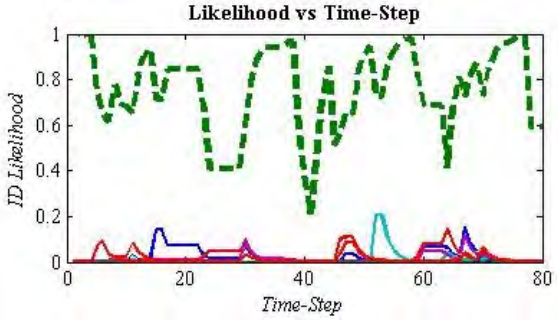

Fig 4. Recognition results fluctuate rapidly without smoothing.

likelihood score of 0.95 in the first frame will incorrectly identify the individual for the entire interaction.

To overcome such bias, each identity track preserves the best speaker likelihood, P_{best_prob}, and identity P_{bestID}. P_{best_prob} is decayed by a constant multiplier (α=0.98) every time-step, and updated when the maximum fusion score is greater than the current speaker likelihood (Eq 5).

$$P_{best\_prob,i}(t) = \max\left[\alpha \cdot P_{best\_prob,i}(t-1), \max\left(P_{prob,i}(t)\right)\right] \quad (5)$$

The identify of each identity track i at time t, therefore, is $P_{bestID,i}(t)$, with likelihood $P_{best\_prob,i}(t)$. When the identity score is greater than some threshold ($P_{best\_prob}(t)>0.7$), the speech data is identified as belonging to $P_{bestID}(t)$, and their speaker model can be updated. To reduce computational load, all active speaker models were recalculated every 30s instead of after every recently identified sample.

V. Performance Evaluation

Performance of the combined system is evaluated in two trials. In the first trial, participants talked to the robot in a 1:1 interaction. These interactions were evaluated to determine classification performance using variations of the learning system described above, as well the resulting speaker recognition performance when the models are retrained with the new data. The second interaction then focused on multi-speaker interactions in the robot's vicinity. The classification performance of the system was evaluated when people were not necessarily talking to the robot.

A. Learning in 1:1 Interactions

To evaluate learning in 1:1 interactions, we used recorded data from a previous person identification trial [7] in which 10 speakers interacted with the robot individually. The interaction consisted of the robot requesting that a user speak on a topic for ~30-sec, and recording the result. 8 different free speech segments per person, averaging 20-sec of speech per speaker, were recorded in this fashion. To vary interaction positions and distances, speakers were directed by the experimenter to stand in four different locations. All speakers followed the same pattern. Because the data used for this evaluation predates the implementation, the robot did not rotate in response to sound localization, and the audio-video data had to be synchronized by hand for each speaker.

Of the 8 recorded free speech segments, 4 were set aside for training and testing. These were split into 2-sec frames, 15 per speaker for training and 24 for testing. The remaining 4 free speech segments were used to learn speaker recognition models autonomously. Given some or no a priori knowledge of individual speaker models, the system used the visual and/or speech data to identify the speaker and associate speech segments. Table 1 describes the classification performance of the system with respect to the set reserved for autonomous recognition. Samples recognized here were then used to update speaker models.

In general, the classification performance of the system looks pretty strong in a 1:1 interaction. 60-70% of the speech vectors are correctly classified and added to the proper speaker models (true positive rate). Furthermore false positive rates are very low, and precision is high with and without supplemental speaker knowledge. Without the weak speaker models created from limited training data, the false positive rate was higher, but still around 1%. As demonstrated in Table 2, however, the greater false positive rate does not result in lower speaker recognition precision.

Table 1. Classification performance during 1:1 interactions.

	True Pos	False Pos	Precision
Face Only	66.1%	1.1%	98.3%
Face + Speaker (All Models)	60.7%	0.5%	99.1%
Face + Speaker (Dynamic Set)	63.8%	0.3%	99.6%

Table 2. Speaker recognition performance after learning.

	3 Active Speakers		10 Active Speakers	
	Precision	Recall	Precision	Recall
No Learning	93.4	52.7	74.4	26.2
Face Only*	51.9 (94.4)	25.4 (65.5)	17.6 (75.2)	2.6 (40.5)
Face + Speaker (All Models)	94.3	63.4	73.7	37.5
Face + Speaker (Dynamic Set)	94.2	64.4	73.8	41.4

* For Face only trials, speaker recognition models were created with and without (in parens) additional training data

The updated speaker recognition model results demonstrate the strengths and weaknesses of online learning. When only face recognition is used to recognize the speaker, and no additional training data is added to the resulting speaker models, precision and recall rates are much less than the baseline condition (e.g. 'No Learning'). This happens because one speaker is recognized very poorly. Their resulting model is created from <2-sec of recorded speech. Worse yet, all of the mistakes made by the system, are made in recognizing that person as another particular individual, whose model is then negatively impacted.

With prior training data, however, the learning phase, with face only, or face + speaker, positively impacts performance. Precision stays about the same for all systems, but recall increases strongly from the baseline performance. As expected, performance with 3 active speakers is much better than 10 active speakers, which favors a dynamic speaker set model. For these 1:1 interactions, however, the lower false positive classification rate (Table 1) due to the dynamic speaker set did not translate into a significant classification performance gain. More likely, as also evidenced by the performance of Face Only with respect to Face + Speaker (Dynamic), the inclusion of more correct speech samples to the model outweighed the small changes in error.

B. Learning in 2:1 Interactions

For multi-speaker scenarios, we evaluated the implemented system with 2 active speakers in the vicinity of the robot. The speakers were not prompted by the robot at all, but rather conversed with each other in the robot's vicinity. The robot then rotated to face each speaker as they talked. 5 trials were completed with an average of 111 sec of recorded data per session. Each trial consisted of 5 separate speech segments dominated by a single speaker.

A problem that quickly arose during multi-speaker interaction was the orientation of the speakers. Because this

Table 3. Classification performance during 2:1 interactions.

	True Pos	False Pos	Precision
Face Only	0.1%	0%	100%
Face + Speaker (All Models)	48.5%	6.5%	88.1%
Face + Speaker* (Dynamic Set)	34.8% (51.2%)	18.1% (4.3%)	65.7% (92.9%)

*Values in () are from seeding correct speaker models.

was not a true interaction with the robot, but rather an interaction in the robot's vicinity, speakers were facing each other most of the time instead of facing the robot. Face recognition, however, performs optimally when the speaker is facing the camera. The side views meant an increase in false positives from face recognition. As a result, the recall rate when using a vision only algorithm dropped to 0.1%. Virtually nothing was recorded.

By incorporating audio information, however, both recall and precision were greatly improved. Loading all speaker models, some of which were trained across multiple days, precision was increased to 88% and recall to 48%. This is lower than in the 1:1 trials because of the reliance of the system on speaker recognition, which has a lower precision than face recognition. With dynamic set loading, the improvement is not as great. The false positive rate, in particular, is very high. The problem is again the low performance of face recognition with indirectly oriented speakers. Because the speakers were not facing the robot, their speaker models were not loaded in a timely fashion. Furthermore, more erroneous speakers were included in the results. This explanation is verified by seeding the dynamic set with the correct speakers. With the correct speakers included from the beginning, precision reaches 92.9% and recall is 51.2%, better than both other approaches.

VI. Discussion

The goal of this work is to ultimately remove the burden of training and re-training of a robot's person identification system from the human. In this paper, we demonstrated how the updating of speaker recognition models can be accomplished autonomously, and verified the results. With existing face and speaker models in the system, the robot can update its speaker recognition models in 1:1 interactions, not only maintaining precision, but improving recall with training. This means that individuals who interact with the robot can improve their speaker recognition models while performing whatever tasks they needed to perform.

Trials of the system in multi-speaker interactions are also promising. When supplied existing audio models of all potential speakers, precision is 88% and recall is 48%. Although such precision is probably too low for learning, because recall is also high, we can use post-processing to raise precision at the expense of recall. A simple example is to only include new audio data from identity tracks whose identity never changes throughout the interaction. This technique raises precision in the multi-speaker tests to 100%, while lowering recall to 9.6%. Other forms of post-processing may raise recall rates even higher. Still, even a low recall value may be fine for such interactions, because, assuming recent speaker models already exist, the robot does not need a lot of additional data to keep existing models updated. Furthermore, simply allowing the robot to maintain its models from other conversations may give it access to a much greater sample set then otherwise available. Then, high precision but small recall, would provide plenty of new data.

Where the online learning of speaker models still needs significant work is in acquiring initial speaker recognition models. The system currently depends too much on a single classifier (i.e. face recognition), which fails with indirect speaker orientation. This dependence similarly limits dynamic speaker sets, which boosted speaker recognition performance when the right people were loaded. The answer to this problem is to incorporate more person identification criteria with different failure points. If face recognition fails, then maybe shirt recognition is a viable alternative.

VII. References

[1] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399-458, Dec 2003.

[2] F. Krsmanovic, C. Spencer, D. Jurafsky, and A. Ng, "Have we met? MDP Based Speaker ID for Robot Dialogue," in *9th International Conf. on Spoken Language Processing*, 2006.

[3] T. Quatiri, *Discrete Time Speech Signal Processing*. Delhi, India: Pearson Education Inc, 2002.

[4] M.S. Nixon and J. Carter, "Automatic Recognition By Gait," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2013 - 2024, 2006.

[5] N. Bellotto and H. Hu, "Multimodal Perception and Recognition of Humans with a Mobile Service Robot," in *International Conference on Robotics and Biomimetics*, 2008.

[6] S. Palanivel and B. Yegnanarayana, "Multimodal person authentication using speech, face and visual speech," *Computer Vision and Image Understanding* , vol. 109, no. 1, pp. 44-55, 2008.

[7] E Martinson, W. Lawson, and G. Trafton, "Person Identification through Perceptual Fusion," IEEE Int Conf on Humanoid Robots, Nashville, TN, 2010.

[8] A. Salah et al., "Multimodal identification and localization of users in a smart environment," *Journal of Multimodal User Interfaces*, vol. 2, pp. 75-91, 2008.

[9] A. Roy and S. Marcel, "Introducing Crossmodal Biometrics:Person Identification from Distinct Audio & Visual Streams," in *IEEE Fourth International Conference on Biometrics: Theory, Applications and Systems*, 2010.

[10] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Towards Development of a Face Recognition System for Watch-List Surveillance," to be published in *IEEE Trans on Pattern Analysis and Machine Intelligence,* 2011.

[11] B Fransen et al., "Using vision, acoustics, and natural language for disambiguation," in *ACM/IEEE International Conference on Human-Robot Interaction*, Arlington, VA, 2007, pp. 73 - 80.

[12] E. Martinson and A. Schultz, "Discovery of sound sources by an autonomous mobile robot," *Autonomous Robots*, vol. 27, no. 3, pp. 221-237, 2009